*Statistical Science*
2015, Vol. 30, No. 3, 352–371
DOI: 10.1214/15-STS513


# A Closer Look at Testing the "No-Treatment-Effect" Hypothesis in a Comparative Experiment

**Joseph B. Lang**

*Abstract.* Standard tests of the "no-treatment-effect" hypothesis for a comparative experiment include permutation tests, the Wilcoxon rank sum test, two-sample $t$ tests, and Fisher-type randomization tests. Practitioners are aware that these procedures test different no-effect hypotheses and are based on different modeling assumptions. However, this awareness is not always, or even usually, accompanied by a clear understanding or appreciation of these differences. Borrowing from the rich literatures on causality and finite-population sampling theory, this paper develops a modeling framework that affords answers to several important questions, including: exactly what hypothesis is being tested, what model assumptions are being made, and are there other, perhaps better, approaches to testing a no-effect hypothesis? The framework lends itself to clear descriptions of three main inference approaches: process-based, randomization-based, and selection-based. It also promotes careful consideration of model assumptions and targets of inference, and highlights the importance of randomization. Along the way, Fisher-type randomization tests are compared to permutation tests and a less well-known Neyman-type randomization test. A simulation study compares the operating characteristics of the Neyman-type randomization test to those of the other more familiar tests.

*Key words and phrases:* Causal effects, completely randomized design, finite-population sampling theory, Fisher vs. Neyman, Fisher's exact test, Horvitz–Thompson estimator, nonmeasurable probability sample, permutation tests, potential variables, process-based inference, randomization-based inference, randomization tests, selection-based inference.

## 1. INTRODUCTION

We begin with a simple example of a randomized comparative experiment. Researchers are interested in determining whether cell phone use while driving has an impact on reaction times. Toward this end, 64 University of Utah student volunteers were enlisted to take part in a randomized comparative experiment (Strayer and Johnston, 2001). Of the 64 students, 32 were randomized to treatment 1 (operate a driving simulator while using a cell phone) and 32 were randomized to treatment 2 (operate a driving simulator without a cell phone). For a summary description of the data and of the way the two treatments were actually administered, see Agresti and Franklin (2007, page 446). In the driving simulation, each student encountered several red lights at random times. Each student's response was the average time required to stop when a red light was detected. The 64 responses, in milliseconds, are recorded in Table 1.

Is there a cell phone use effect? Generically, is there a *treatment effect*?

Standard tests of the "no-treatment-effect" hypothesis include permutation tests (Pitman, 1937, 1938), the Wilcoxon rank sum test (Wilcoxon, 1945), two-sample $t$ tests (cf. Welch, 1938), and Fisher-type ran-

*Joseph B. Lang is Professor, Department of Statistics and Actuarial Science, University of Iowa, 207 SH, Iowa City, Iowa 52242, USA (e-mail: joseph-lang@uiowa.edu).*

TABLE 1
*Reaction times* (*milliseconds*)

| | |
|---|---|
| Cell phone: | 636 623 615 672 601 600 542 554 543 520 609 559 595 565 573 554 |
| | 626 501 574 468 578 560 525 647 456 688 679 960 558 482 527 536 |
| Control: | 557 572 457 489 532 506 648 485 610 444 626 626 426 585 487 436 |
| | 642 476 586 565 617 528 578 472 485 539 523 479 535 603 512 449 |
| Generically... | |
| Treatment 1: | $y_{1,1}, y_{1,2}, \ldots, y_{1,32}$ |
| Treatment 2: | $y_{2,1}, y_{2,2}, \ldots, y_{2,32}$. |

domization tests (Eden and Yates, 1933, Fisher, 1935; see also the history in David, 2008). Most practitioners are aware that these procedures test different "no-effect" hypotheses and are based on different modeling assumptions. However, this awareness is not always, or even usually, accompanied by a clear understanding or appreciation of these differences. This paper looks at each of these testing approaches and addresses the all important questions, exactly what hypothesis is being tested and what model assumptions are being made? Along the way, we will have to confront several other questions such as, how is the definition of *treatment effect* operationalized, what is the actual target of inference, what is the role of randomization, and are there other, perhaps better, approaches to testing a no-effect hypothesis?

To address these questions, we draw on ideas from the rich literature on causal analysis. In particular, we employ the useful concept of "potential variables." Although the idea of potential variables can be traced back to Neyman (1923), Rubin, beginning with a series of papers on causal models in the 1970s (see Rubin, 2010, and references therein) is usually credited with more explicitly stating the potential variable model and extending it to both randomized and nonrandomized design settings, with or without covariates (see Rubin's causal model, Holland, 1986). Between Neyman and Rubin, potential variables were used by relatively few authors; Welch (1937), Kempthorne (1952, 1955), and Cox (1958a), Section 5, were among the notable early proponents. Around the time of and after Rubin, many more authors made important contributions to the potential variables literature. See, for example, Copas (1973), Bailey (1981), Holland (1986), Greenland (1991, 2000), Gadbury (2001), and the references therein.

To be clear, it is not the goal of this paper to summarize the vast literature on potential variables and causal modeling. (To this end, see Paul R. Rosenbaum's very informative website and references therein, www-stat.wharton.upenn/~rosenbap/downloadTalks.htm.)

Instead, the first goal is to exploit the benefits of hindsight to develop a modeling framework that supports clear descriptions and comparisons of the different testing approaches, and promotes careful consideration of the model assumptions and targets of inference. This modeling framework and associated notation draws clear distinctions between realizations and random variables, and between observed and unobserved data. It accommodates both treatment assignment and sampling from populations, and clearly differentiates between the two. Although the proposed model lends itself to generalizations in many directions (e.g., more than two treatments, restricted randomization, etc.), to simplify exposition, we will focus on the two-treatment comparative experiment setting. This restriction allows us to more directly highlight the useful features of the proposed modeling framework.

The second goal of this paper is to address the question of availability of other testing approaches, besides the four common ones mentioned above. Toward this end, we revisit ideas introduced in Neyman (1923). Using the model structure introduced herein, we describe a less well-known Neyman-type randomization test, which is qualitatively different than the Fisher-type randomization test (cf. Welch, 1937; Rubin, 1990, 2004, 2010). (Readers with an interest in history are encouraged to read Neyman et al., 1935, along with the discussions, to see how Neyman and Fisher publicly aired their differences of opinions on testing in randomized design settings.) The Neyman randomization test, which uses a less restrictive "no-effect" hypothesis than Fisher's, is based on a test statistic with the common form, (estimator minus estimand)/(standard error of estimator). Neyman, with an eye on interval estimation rather than testing, derived the standard error with respect to a randomization distribution using tools from finite-population sampling theory. In retrospect, Neyman's derivation approach is hardly surprising given that he "may be said to have initiated the modern theory of survey sampling" (Lehmann, 1994) in his landmark paper of 1934 (Neyman, 1934). Compared to Fisher

randomization tests, the Neyman tests do have their advantages and disadvantages. One disadvantage is that Neyman tests are approximate, whereas Fisher tests are exact. An advantage is that the Neyman test can be more powerful than the Fisher test (see Section 7 below). Another advantage is that, unlike the Fisher randomization test, the Neyman version can be used to test hypotheses about a population when units are randomly sampled from the population and then randomized to treatment levels.

The third and final goal of this paper is to compare the operating characteristics of the five tests: the permutation test, the Wilcoxon rank sum test, the two-sample $t$-test, the Fisher randomization test, and the Neyman randomization test. The penultimate section of the paper includes a small-scale simulation study of the size and power of these five tests. Based on these comparisons, we make tentative recommendations on which test to use in different settings.

The remainder of this paper is organized as follows: Section 2 introduces potential variables and recasts the data in Table 1 within this framework. Section 3 introduces a sequential data generation model that explicitly accommodates both random sampling and randomization. The components in the three-level sequential model are identified as the "process," the "sampling," and the "randomization." This model, along with a useful component-selection notation, leads to an explicit identification of the observed data and the three main targets of inference. Section 4 gives candidate definitions of treatment effects that are based on potential variables; corresponding no-treatment-effect hypotheses are also given. An overview of the three main inference approaches—process-based, selection-based, and randomization-based—is given in Section 5. Section 6 introduces a difference statistic that can be used as the basis for tests of the no-treatment-effect hypotheses. Tests corresponding to each of the three inference approaches, along with assumptions for their validity, are described in detail; some of these tests are well known and some are less well known. Section 7 carries out an analysis of the cell phone data and includes a small-scale simulation study of the operating characteristics of the different testing approaches discussed herein. Finally, Section 8 includes a brief discussion.

## 2. WHAT MIGHT HAVE BEEN: THE POTENTIAL VARIABLES VIEWPOINT

Going back to Neyman (1923) and following the lead of Welch (1937), Kempthorne (e.g., 1955, 1977), Cox (1958a), and Rubin (e.g., 2005), we will view the data as observed values of a sample of "potential values."

Consider a population of $N$ units that are, without loss of generality, identified by the numbers 1 through $N$; in symbols, we will let $\underline{P} = (1, \ldots, N)$ represent the unit identifiers for the population. For convenience, we will also refer to $\underline{P}$ as the population of units. Let $Y_{t.i}$ be the response for unit $i$ when exposed to treatment $t$, where $i = 1, \ldots, N$ and $t = 1, 2$. The response variables $Y_{1.i}$ and $Y_{2.i}$ are called potential variables for reasons made clear in the next paragraph.

The introduction of these potential variables leads to intuitively appealing definitions of treatment effects that are based on head-to-head comparisons of $Y_{1.i}$ and $Y_{2.i}$. There is a catch, however. Although there is the potential to observe either $Y_{1.i}$ or $Y_{2.i}$, unfortunately, it is not possible to observe both. Strictly speaking, it is not possible to observe the values of both potential variables because the same subject cannot be simultaneously exposed to both treatments. To the potential variable advocates, this is the "fundamental problem of causal analysis" (Holland, 1986). As an example, if we observe the value of $Y_{2.i}$, then the value of $Y_{1.i}$, and hence the difference $Y_{1.i} - Y_{2.i}$, cannot be observed. In this case, the unobserved value of $Y_{1.i}$ is relegated to counterfactual status; the value is "what might have been" had unit $i$ been exposed to treatment 1 rather than treatment 2.

The data in Table 1 can be viewed as observed values of a sample of the potential variable values. Specifically, a sample $\underline{s}$ of size $n = n_1 + n_2 = 64$ is taken, without replacement, from the population $\underline{P}$. That is, $\underline{s} = (s_1, \ldots, s_n)$, where $s_j \in \underline{P}$ and $s_j \neq s_{j'}$. One of the two treatments will be assigned to each of the units in the sample $\underline{s}$. For the example, treatment 1 was assigned to $n_1 = 32$ and treatment 2 was assigned to $n_2 = 32$ of the 64 sampled units.

Let $y_{t.s_j}$ be the response value for sampled subject $s_j$ when exposed to treatment $t$. That is, $y_{t.s_j}$ is a realization of $Y_{t.s_j}$. Of course, for each subject $s_j$, only one of the realizations, $y_{1.s_j}$ or $y_{2.s_j}$, will be observed. From a potential variables viewpoint, the original data in Table 1 can be viewed as follows:

REMARK. Table 1 used the conventional $y_{t,i}, i = 1, \ldots, 32, t = 1, 2$ to represent the observed data, whereas Table 2 uses $y_{2.s_1}, y_{2.s_2}, y_{1.s_3}, \ldots, y_{2.s_{63}}, y_{1.s_{64}}$. It is important to note that the symbols $y_{t,i}$ and $y_{t.i}$ represent very different objects. For example, in Table 1, of those units sampled and assigned treatment 1, the

TABLE 2
*Potential values notation*

| | | |
|---|---|---|
| Treatment 1: | $\cancel{y_{1.s_1}}\ \cancel{y_{1.s_2}}\ y_{1.s_3}, \ldots, y_{1.s_{63}}\ \cancel{y_{1.s_{64}}}$ | Only the 32 non-$\times$'ed out values are observed. |
| Treatment 2: | $y_{2.s_1}, y_{2.s_2}, \cancel{y_{2.s_3}}, \ldots, y_{2.s_{63}}, \cancel{y_{2.s_{64}}}$ | Only the 32 non-$\times$'ed out values are observed. |

Here, $\underline{s} = (s_1, \ldots, s_{64})$ is a sample from some population $\underline{P} = (1, \ldots, N)$, $N \geq 64$.

3rd had a process value of $y_{1,3}$, and of those units sampled and assigned treatment 2, the 3rd had a process value of $y_{2,3}$. That is, $y_{1,3}$ and $y_{2,3}$ are process values for two *distinct units*. In contrast, $y_{1.3}$ and $y_{2.3}$ represent the process values under treatments 1 and 2 for the *same unit*, specifically, the 3rd unit in the population.

## 3. DATA-GENERATION MODELS AND INFERENCE GOALS

Let $\underline{Y} = (Y_{1.1}, \ldots, Y_{1.N}, Y_{2.1}, \ldots, Y_{2.N})$ be the vector of potential variables for the population $\underline{P}$ and $\underline{y} = (y_{1.1}, y_{1.2}, \ldots, y_{1.N}, y_{2.1}, \ldots, y_{2.N})$ be the corresponding vector of realizations. We will use this notational convention throughout the paper: upper case letters for random variables and lower case letters for realizations.

To simplify and to highlight vector component identification, we introduce dot "." operations and a component-selection bracket "[ ]" notation that is similar to the matrix syntax used in computer languages such as R. Let $\underline{x}$ and $\underline{w}$ be $m$-dimensional vectors and let $k$ be a scalar. Define

$$\underline{x}.\underline{w} = (x_1.w_1, \ldots, x_m.w_m) \quad \text{and}$$
$$k.\underline{x} = (k.x_1, \ldots, k.x_m).$$

Consider an $m$-dimensional vector $\underline{x}$ with components identified by subscripts $a_1, \ldots, a_m$, that is, $\underline{x} = (x_{a_1}, \ldots, x_{a_m})$. Provided $\underline{b} = (b_1, \ldots, b_q)$ has components $b_i \in \{a_1, \ldots, a_m\}$, for each $i = 1, \ldots, q$, the vector $\underline{x}[\underline{b}]$ is defined as $\underline{x}[\underline{b}] = \underline{x}[b_1, \ldots, b_q] = (x_{b_1}, \ldots, x_{b_q})$.

As an example, $\underline{y} = (y_{1.1}, \ldots, y_{1.N}, y_{2.1}, \ldots, y_{2.N})$ can be expressed as $\underline{y} = \underline{y}[1.\underline{P}, 2.\underline{P}]$. Similarly, $\underline{y}[1.\underline{s}] = (y_{1.s_1}, \ldots, y_{1.s_n})$ and $\underline{y}[\underline{t}.\underline{s}] = (y_{t_1.s_1}, \ldots, y_{t_n.s_n})$. We will also use a notation for averages: As examples,

$$\overline{Y}[t.\underline{P}] = N^{-1} \sum_{i=1}^{N} \underline{Y}[t.i],$$

$$\overline{y}[t.\underline{P}] = N^{-1} \sum_{i=1}^{N} \underline{y}[t.i] \quad \text{and}$$

$$\overline{y}[t.\underline{s}] = n^{-1} \sum_{j=1}^{n} \underline{y}[t.s_j].$$

The data-generation models we consider in this paper are based on the following sequential generations:

$$\begin{aligned} \underline{y} &\leftarrow \underline{Y} \\ &\text{here, } \underline{y} = (y_{1.1}, \ldots, y_{1.N}, y_{2.1}, \ldots, y_{2.N}), \\ \underline{s} &\leftarrow \underline{S} | (\underline{Y} = \underline{y}) \\ &\text{here, } \underline{s} = (s_1, \ldots, s_n), s_j \in \underline{P}, s_j \neq s_{j'}, \\ \underline{t} &\leftarrow \underline{T} | (\underline{Y} = \underline{y}, \underline{S} = \underline{s}) \\ &\text{here, } \underline{t} = (t_1, \ldots, t_n), t_j \in \{1, 2\}. \end{aligned} \tag{1}$$

The left arrow "$\leftarrow$" is read, "is a realization of." Here, $\underline{Y}$ is the collection of $2N$ potential response variables, $\underline{S}$ is the collection of $n$ sampling variables, and $\underline{T}$ is the collection of $n$ treatment assignment variables. The sequencing in (1) is not required to correspond to the temporal sequencing of data generation. It is meant only as a device for specifying the joint distribution of $(\underline{Y}, \underline{S}, \underline{T})$. For a related discussion, see Rubin [2010, between equations (4) and (5)].

In words, "Nature" generates $N$ units, which are labeled $1, 2, \ldots, N$. Each unit can potentially experience either of two "possible worlds," which correspond to exposure under the two treatments. The vector $\underline{y}$ contains the $2N$ potential response values, one for each of the $N$ units under treatment 1 and one for each of the $N$ units under treatment 2. These potential deviates in $\underline{y}$ are viewed as realized, at least in theory, but only partially observable. We sample $n$ distinct subjects $\underline{s}$ from the population $\underline{P}$. The sampling may depend on potential deviates $\underline{y}$; this dependence often stems from selecting on covariates that are statistically related to the potential variables [see Rubin, 2010, between equations (4) and (5)]. Finally, we assign treatment levels $\underline{t}$ to units in the sample, that is, we choose which of the two possible worlds we will observe for each unit in the sample. The treatment assignment may depend on the potential deviates $\underline{y}$ and/or the sampled units $\underline{s}$. However, when mechanical or physical randomization (cf. Fisher, 1935,

Kempthorne, 1955) is used, the treatment assignment can be made to be independent of the potential deviates.

We will refer to the potential variables $\underline{Y}$ as "process variables"[1] and the values $\underline{y}$ as "process values," to differentiate them from the "selection" variables ($\underline{S}, \underline{T}$) and values ($\underline{s}, \underline{t}$). The process portion describes how things behave under both possible worlds and the selection portion determines how we go about observing this behavior. Owing to the sampling and treatment assignment (the selection), we do not observe the entire vector of potential deviates $\underline{y}$ (the process values). Indeed, the "fundamental problem of causal inference" rules out the possibility of fully observing the $2N$-dimensional data vector $\underline{y}$. Instead, we observe only the $n$-dimensional sub-vector $\underline{y}[\underline{t}.\underline{s}]$. Schematically, we have

$$\underbrace{\underline{y}[\underline{t}.\underline{s}]}_{\text{observed}} \subseteq \underbrace{\underline{y}[1.\underline{s}, 2.\underline{s}] \subseteq \underline{y}[1.\underline{P}, 2.\underline{P}] = \underline{y} \leftarrow \underline{Y}}_{\text{unobserved}}.$$

The inference goal of this paper can be stated succinctly as follows:

*Inference goal*. Use the observed data $\underline{y}[\underline{t}.\underline{s}]$ from a comparative experiment to reduce uncertainty about one of the three targets: the vector $\underline{y}[1.\underline{s}, 2.\underline{s}]$, the vector $\underline{y}[1.\underline{P}, 2.\underline{P}]$, or the distribution of $\underline{Y}$.

## 4. TREATMENT EFFECTS AND "NO-TREATMENT-EFFECT" HYPOTHESES

### 4.1 Treatment Effects

We began this paper with the question of whether there was a treatment effect. Of course, this raises another question: What exactly is a "treatment effect"?

In a comparative experiment, a treatment effect can be viewed as some measure of the difference between the response ($\underline{Y}$) distribution or response values ($\underline{y}$) for treatment level 1 and the response distribution or response values for treatment level 2. The potential variables viewpoint lends itself to intuitively-appealing candidate definitions of such treatment effects (cf. Neyman, 1923; Rubin, 1990, 2005, 2010). Some of the candidates considered in this paper are as follows:

Realized unit-specific effects:
$\underline{y}[1.s_j] - \underline{y}[2.s_j], j = 1, \ldots, n$ or
$\underline{y}[1.i] - \underline{y}[2.i], i = 1, \ldots, N.$
Distribution unit-specific effects:
$\delta(F_{1.i}, F_{2.i}), i = 1, \ldots, N.$
Expected unit-specific effects:
$E(\underline{Y}[1.i]) - E(\underline{Y}[2.i]), i = 1, \ldots, N.$
Realized aggregate effects:
$\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$ or
$\overline{y}[1.\underline{P}] - \overline{y}[2.\underline{P}].$
Expected aggregate effects:
$E(\overline{Y}[1.\underline{P}]) - E(\overline{Y}[2.\underline{P}]).$

For example, the realized unit-specific treatment effect $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ is simply the difference between unit $s_j$'s responses under two scenarios or two possible worlds—in one world the unit is exposed to treatment 1 and in the other world the unit is exposed to treatment 2. As another example, the distribution unit-specific effect $\delta(F_{1.i}, F_{2.i})$ measures the distance between the c.d.f.'s of $\underline{Y}[1.i]$ and $\underline{Y}[2.i]$ using some distance function $\delta(\cdot)$. This latter example illustrates that treatment effects need not be defined in terms of simple differences, arithmetic averages, or means of distributions. Other examples of treatment effects include the distribution unit-specific effect median($\underline{Y}[1.i]$) − median($\underline{Y}[2.i]$), realized unit-specific effects, such as $(\underline{y}[2.s_j] - \underline{y}[1.s_j])/\underline{y}[1.s_j]$, and realized aggregate effects, such as $\|\overline{y}[1.\underline{s}] - \underline{y}[2.\underline{s}]\|$ or var($\underline{y}[1.\underline{s}]$) − var($\underline{y}[2.\underline{s}]$) or $\frac{\overline{y}[2.\underline{s}]-\overline{y}[1.\underline{s}]}{\overline{y}[1.\underline{s}]}$, or for binary responses, the realized odds ratio $\frac{\overline{y}[1.\underline{s}]/(1-\overline{y}[1.\underline{s}])}{\overline{y}[2.\underline{s}]/(1-\overline{y}[2.\underline{s}])}$, etc.

Unfortunately, none of the treatment effects mentioned above is observable. The expected and distribution effects cannot be observed because the distribution of $\underline{Y}$ is not completely known. The realized effects cannot be observed because, by the fundamental problem of causal inference, only one of the realizations, for example, either $\underline{y}[1.s_j]$ or $\underline{y}[2.s_j]$, can be observed. Fortunately, this does not preclude unbiased estimation of some of these unobservable treatment effects, as we point out below.

In the potential-variables causal literature, the treatment effects defined above would be considered causal effects provided certain assumptions hold (e.g., Rubin, 1990, 2005, 2010). To avoid the ongoing debate about the nature of causality, we will refrain from referring to treatment effects as causal effects.

### 4.2 "No-Treatment-Effect" Hypotheses

Corresponding to each treatment effect definition, there is a "no-treatment-effect" hypothesis. As exam-

[1] Rubin (2005), uses the descriptor "science" rather than "process."

ples,

$$H_0^{UP}: \underline{Y}[1.\underline{P}] = \underline{Y}[2.\underline{P}], \quad \text{with probability 1;}$$

$$H_0^{DUP}: \underline{Y}[1.i] \sim \underline{Y}[2.i], i = 1, \ldots, N.$$

Herein, "$\sim$" means "distributed as";

$$H_0^{EUP}: E(\underline{Y}[1.i]) = E(\underline{Y}[2.i]), \quad i = 1, \ldots, N;$$

$$H_0^{RUP}: \underline{y}[1.\underline{P}] = \underline{y}[2.\underline{P}];$$

$$H_0^{RAP}: \overline{y}[1.\underline{P}] = \overline{y}[2.\underline{P}];$$

$$H_0^{RUs}: \underline{y}[1.\underline{s}] = \underline{y}[2.\underline{s}];$$

$$H_0^{RAs}: \overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}].$$

The indentations are used to denote nesting. For example, both $H_0^{EUP}$ and $H_0^{RUP}$ are implied by $H_0^{UP}$. Similarly, $H_0^{RAs}$ is implied by $H_0^{RUs}$ and by $H_0^{RUP}$, but not by $H_0^{RAP}$. The superscripts remind us of the type of treatment effect used in the hypothesis. For example, the hypothesis $H_0^{EUP}$ uses *E*xpected *U*nit-specific effects (for the *P*opulation), and $H_0^{RAs}$ uses *R*ealized *A*ggregate (over *s*ample $\underline{s}$) effects.

## 5. INFERENCE APPROACHES AND ASSUMPTIONS

The $(\underline{y}, \underline{s}, \underline{t})$ components in the observed data $\underline{y}[\underline{t}.\underline{s}]$ are viewed as outcomes of the sequential generations of (1). The complete, but only partially observed, data $\underline{y}$ is a realization of the $2N$-dimensional vector of potential variables $\underline{Y}$. In symbols, we have $\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]$ and $\underline{y} \leftarrow \underline{Y}$.

As stated previously, the inference goal is to use the observed data $\underline{y}[\underline{t}.\underline{s}]$ to reduce uncertainty about one of three targets: the distribution of $\underline{Y}$, the vector $\underline{y}[1.\underline{P}, 2.\underline{P}]$, or the vector $\underline{y}[1.\underline{s}, 2.\underline{s}]$. The choice of inference approach depends on which of these targets we are interested in and it depends on what assumptions we can reasonably make about the joint distribution of $(\underline{Y}, \underline{S}, \underline{T})$, where $\underline{Y}$ is the "process" variable and $(\underline{S}, \underline{T})$ are the "selection" variables. More specifically, $\underline{S}$ is the "sampling" variable and $\underline{T}$ is the "randomization" or treatment assignment variable. In this paper, we consider three candidate inference approaches.

### 5.1 Process-Based Inference and Assumptions

With the process-based approach, we condition on the selection (only $\underline{Y}$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]|(\underline{S} = \underline{s}, \underline{T} = \underline{t}) \sim \underline{Y}[\underline{t}.\underline{s}]|(\underline{S} = \underline{s}, \underline{T} = \underline{t})$$

to carry out inferences about the distribution of $\underline{Y}$. (The discussion section describes more general inferences.)

With process-based inference, we generally must make assumptions about the conditional distribution of $\underline{Y}|(\underline{S} = \underline{s}, \underline{T} = \underline{t})$. However, because this paper focuses on test procedures that are valid provided the process is independent of the selection (see assumption $A_1$), we will only make assumptions about the (unconditional) process distribution of $\underline{Y}$; see assumptions $A_2$–$A_7$.

$A_1: (\underline{S}, \underline{T}) \perp\!\!\!\perp \underline{Y}$;

$A_2: \underline{Y}[t.i] \sim F_t, \quad i = 1, \ldots, N, t = 1, 2$;

$A_3: \underline{Y}[1.i, 2.i], \quad i = 1, \ldots, N,$ are independent;

$A_4: F_t \in \{\text{continuous c.d.f.s}\}$;

$A_5: F_t \in \{N(\mu_t, \sigma_t^2) \text{ c.d.f.s}\}$;

$A_6: F_t \in \{N(\mu_t, \sigma^2) \text{ c.d.f.s}\}$;

$A_7: F_t \in \{\text{c.d.f.s with mean and variance } (\mu_t, \sigma_t^2)\}$.

Process-based inference is simplified under Assumption $A_1$ because the observed data can be viewed as a realization of $\underline{Y}[\underline{t}.\underline{s}]$; in symbols, $\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{t}.\underline{s}]$. It follows that we need only model the (unconditional) distribution of the process variable $\underline{Y}$. Importantly, under $A_1$, process-based inference does not require any assumptions about the selection $(\underline{S}, \underline{T})$. It is also important to note that when $A_1$ holds and $\underline{Y}$ is modeled via assumptions such as $A_2$–$A_7$, the observed data $\underline{y}[\underline{t}.\underline{s}]$ can be used to make process-based inferences about the distribution of $\underline{Y}$.

Unfortunately, Assumption $A_1$ is typically not tenable in practice. Notice that $A_1$ is equivalent to the two assumptions, $\underline{T} \perp\!\!\!\perp \underline{Y}|\underline{S}$ and $\underline{S} \perp\!\!\!\perp \underline{Y}$. When mechanical randomization is used to assign treatments to the sampled units, the first assumption can be made tenable. However, the reasonableness of the second assumption, that the sampling variable $\underline{S}$ is independent of $\underline{Y}$, is often questionable in practice. For example, with haphazard or convenience sampling, rather than probability sampling, it often turns out that $\underline{S}$ and $\underline{Y}$ are not independent. The dependence typically stems from sampling on the basis of covariates that are related to $\underline{Y}$.[2]

The assumption $A_2$ is not as restrictive as it may initially appear. For example, whenever the identifiers are

[2]Of course, if the covariates responsible for the dependence were known and observable, we could condition on their values to restore independence; however, this conditional model falls outside the purview of the current paper.

arbitrarily assigned to the $N$ population units, the $N$ pairs $\underline{Y}[1.i, 2.i]$ would be exchangeable and, hence, $A_2$ would hold. Generally, the more tenuous assumptions are $A_1$, that the selection is carried out independently of the process, the independence assumption $A_3$, and the assumptions $A_4$–$A_7$ about the form of marginal distributions $F_1$ and $F_2$.

### 5.2 Randomization-Based Inference and Assumptions

With the randomization-based approach, we condition on both the process values and the sample (only $\underline{T}$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]|(\underline{Y} = \underline{y}, \underline{S} = \underline{s}) \sim \underline{y}[\underline{T}.\underline{s}]|(\underline{Y} = \underline{y}, \underline{S} = \underline{s})$$

to carry out inferences about $\underline{y}[1.\underline{s}, 2.\underline{s}]$.

With randomization-based inference, we generally must make assumptions about the conditional distribution of $\underline{T}|(\underline{Y} = \underline{y}, \underline{S} = \underline{s})$. However, because this paper focuses on test procedures that are valid provided the randomization is conditionally independent of the process, given the sample (see assumption $B_1$), we will only make assumptions about the distribution of $\underline{T}|(\underline{S} = \underline{s})$; see assumption $B_2$.

$B_1 : \underline{T} \perp\!\!\!\perp \underline{Y}|\underline{S}$.

$B_2$ : The distribution of $\underline{T}|(\underline{S} = \underline{s})$ is completely known and satisfies...

$$P(\underline{T}.\underline{S} \ni t.s_j|\underline{S} = \underline{s}) > 0, \quad j = 1, \ldots, n, t = 1, 2;$$

$$P(\underline{T}.\underline{S} \ni t.s_j, \underline{T}.\underline{S} \ni t'.s_{j'}|\underline{S} = \underline{s}) = 0 \quad \text{if and only if } t \neq t' \text{and} j = j'.$$

Randomization-based inference is simplified under assumption $B_1$ and fortunately the use of mechanical randomization makes this assumption tenable. Under $B_1$, we have that the observed data can be viewed as $\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{y}[\underline{T}.\underline{s}]|(\underline{S} = \underline{s})$, so only the distribution of $\underline{T}|(\underline{S} = \underline{s})$ needs to be modeled. In particular, we need not make any assumption about the distribution of $\underline{Y}$ or its relation to $\underline{S}$; for example, $\underline{Y}$ and $\underline{S}$, that is, the process and the sampling, need not be independent. It is important to note that when the distribution of $\underline{T}|(\underline{S} = \underline{s})$ is completely known (see $B_2$), the distribution of $\underline{y}[\underline{T}.\underline{s}]|(\underline{S} = \underline{s})$ is known up to the partially-observed values $\underline{y}[1.\underline{s}, 2.\underline{s}]$, which are the parameters of interest for randomization-based inference. Thus, when $B_1$ and $B_2$ hold, the observed data $\underline{y}[\underline{t}.\underline{s}]$ can be used to carry out randomization-based inference about the target parameters $\underline{y}[1.\underline{s}, 2.\underline{s}]$.

In $B_2$, the probabilities are called first- and second-order inclusion probabilities for the random sample, namely, $\underline{T}.\underline{S}|(\underline{S} = \underline{s})$, taken from $(1.\underline{s}, 2.\underline{s})$. Assumption $B_2$ imposes constraints on these inclusion probabilities. The positive first-order inclusion probabilities imply that "proper" randomization is used to assign treatments, that is, each unit in the sample has a positive probability of receiving either treatment; we say that this is a "*proper*" *randomized comparative experiment*.[3] Put another way, $\underline{T}.\underline{S}|(\underline{S} = \underline{s})$ is a *probability* sample from $(1.\underline{s}, 2.\underline{s})$. Because the same unit cannot be assigned different treatments, the second-order inclusion probabilities with $t \neq t'$ and $j = j'$ are 0. This implies that the probability sample is nonmeasurable, to use language from sampling theory (cf. Särndal et al., 1992, pages 32–33). This nonmeasurability complicates the computation of certain randomization-based test statistics (see Section 6.3.2 below), as Neyman was fully aware of in 1923.

### 5.3 Selection-Based Inference and Assumptions

With the selection-based approach, we condition on the process values [only $(\underline{S}, \underline{T})$ is random] and use

$$\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]|(\underline{Y} = \underline{y}) \sim \underline{y}[\underline{T}.\underline{S}]|(\underline{Y} = \underline{y})$$

to carry out inferences about $\underline{y}[1.\underline{P}, 2.\underline{P}]$.

With selection-based inference, we generally must make assumptions about the conditional distribution of $(\underline{S}, \underline{T})|(\underline{Y} = \underline{y})$. However, because this paper focuses on test procedures that are valid provided the selection is independent of the process (see assumption $C_1$), we will only make assumptions about the unconditional distribution of $(\underline{S}, \underline{T})$; see assumption $C_2$.

$C_1 : (\underline{S}, \underline{T}) \perp\!\!\!\perp \underline{Y}$.

$C_2$ : The distribution of $(\underline{S}, \underline{T})$ is completely known and satisfies...

$$P(\underline{T}.\underline{S} \ni t.i) > 0, \quad i = 1, \ldots, N, t = 1, 2.$$

$$P(\underline{T}.\underline{S} \ni t.i, \underline{T}.\underline{S} \ni t'.i') = 0 \quad \text{if and only if } t \neq t' \text{ and } i = i'.$$

[3] The two-treatment completely randomized design (CRD) experiment is a special-case example of a proper randomized comparative experiment. With the CRD, $\underline{T}|(\underline{S} = \underline{s})$ has a uniform distribution over all possible rearrangments of $n_1$ 1's and $n_2$ 2's (cf. Cox, 1958b, pages 71–72; Kempthorne, 1977, Section 8; or Cox and Reid, 2000, Section 2.2.4.)

Selection-based inference is simplified under assumption $C_1$ because the observed data can be viewed as $\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{y}[\underline{T}.\underline{S}]$. It follows that we need only specify the (unconditional) distribution of the selection $(\underline{S}, \underline{T})$; no assumptions about $\underline{Y}$ are needed. Unfortunately, as discussed in the process-based subsection above, assumption $C_1$ is not usually tenable in practice because the sampling and process are often dependent. It is important to note that when the distribution of $(\underline{S}, \underline{T})$ is completely known (see $C_2$), the distribution of $\underline{y}[\underline{T}.\underline{S}]$ is known up to the partially-observed values $\underline{y}[1.\underline{P}, 2.\underline{P}]$, which are the parameters of interest for selection-based inference. Thus, when $C_1$ and $C_2$ hold, the observed data $\underline{y}[\underline{t}.\underline{s}]$ can be used to carry out selection-based inference about the target parameters $\underline{y}[1.\underline{P}, 2.\underline{P}]$.

As discussed in the randomization-based section, assumption $C_2$ imposes constraints on first- and second-order inclusion probabilities. In this case, the random sample $\underline{T}.\underline{S}$ is taken from $(1.\underline{P}, 2.\underline{P})$. The assumption implies that each of the $2N$ elements in $(1.\underline{P}, 2.\underline{P})$ has a positive probability of being selected. Thus, the random sample is a probability sample. The 0 second-order inclusion probabilities imply that the probability sample is nonmeasurable.

## 6. TESTS OF THE NO-TREATMENT-EFFECT HYPOTHESIS

This section describes a collection of process-, randomization-, and selection-based tests of no treatment effect hypotheses. Some of these tests are well known (e.g., the two-sample $t$ test), and some are less well known (e.g., the Neyman randomization test). In any case, we will emphasize the assumptions needed for their applicability and we will carefully state the hypothesis that is actually being tested. We begin by introducing a difference statistic that forms the basis of most of the tests considered in this paper.

### 6.1 The Difference Statistic

With the exception of the Wilcoxon rank sum statistic, this paper will focus on test statistics that are based on the following difference statistics:

$$\underbrace{D_1(\underline{Y}[\underline{t}.\underline{s}]) = D(\underline{Y}, \underline{s}, \underline{t}, \underline{w}_1)}_{\text{process}},$$

$$\underbrace{D_3(\underline{T}) = D(\underline{y}, \underline{s}, \underline{T}, \underline{w}_3)}_{\text{randomization}},$$

$$\underbrace{D_{23}(\underline{S}, \underline{T}) = D(\underline{y}, \underline{S}, \underline{T}, \underline{w}_{23})}_{\text{selection}},$$

where

$$D(\underline{y}, \underline{s}, \underline{t}, \underline{w}) = \underbrace{\sum_{i=1}^{N} \frac{\underline{y}[1.i]\mathbf{1}(\underline{t}.\underline{s} \ni 1.i)}{\underline{w}[1.i]}}_{\text{weighted avg of trt 1 values}} - \underbrace{\sum_{i=1}^{N} \frac{\underline{y}[2.i]\mathbf{1}(\underline{t}.\underline{s} \ni 2.i)}{\underline{w}[2.i]}}_{\text{weighted avg of trt 2 values}}. \tag{2}$$

The candidate values for weights $\underline{w}$ include

$$\underline{w}_1[t.i] = n_t, \qquad \underline{w}_3[t.i] = nP(\underline{T}.\underline{s} \ni t.i \,|\, \underline{S} = \underline{s}),$$
$$\underline{w}_{23}[t.i] = NP(\underline{T}.\underline{S} \ni t.i),$$

where $n = \text{length}(\underline{s}), n_t = \sum_{j=1}^{n} \mathbf{1}(t_j = t)$. From the discussions in Sections 5.2 and 5.3, it follows that the $\underline{w}_3$ and $\underline{w}_{23}$ components are multiples of first-order inclusion probabilities (cf. Särndal et al., 1992), using language from finite-population sampling theory. By convention, we set $0/0 \equiv 0$ in (2).

There are several useful properties of these $D$ statistics. First, note that

> $D(\underline{y}, \underline{s}, \underline{t}, \underline{w})$ can be computed using only the observed values $\underline{y}[\underline{t}.\underline{s}]$, $\underline{s}$, and $\underline{t}$.

That is, $D$, and hence each of $D_1$, $D_3$, and $D_{23}$, is an observable statistic. It also follows that the process-based statistic $D_1$ depends on $\underline{Y}$ only through $\underline{Y}[\underline{t}.\underline{s}]$, hence the notation $D_1(\underline{Y}[\underline{t}.\underline{s}])$. Second, the process-based statistic $D_1$ is simply the difference between the unweighted sample averages $n_1^{-1}\sum_{j:t_j=1}\underline{Y}[1.s_j]$ and $n_2^{-1}\sum_{j:t_j=2}\underline{Y}[2.s_j]$. The randomization- and the selection-based statistics $D_3$ and $D_{23}$ are differences between probability-weighted sample averages. Third,

> Under $A_1$,$D_1|(\underline{S} = \underline{s}, \underline{T} = \underline{t})$
>
> has distribution that depends only on the model
>
> for $\underline{Y}[\underline{t}.\underline{s}]$.
>
> Under $B_1$,$D_3|(\underline{Y} = \underline{y}, \underline{S} = \underline{s})$
>
> has distribution that depends only on the
>
> $\underline{y}[1.\underline{s}, 2.\underline{s}]$ values and the $\underline{T}|(\underline{S} = \underline{s})$ distribution.
>
> Under $C_1$,$D_{23}|(\underline{Y} = \underline{y})$
>
> has distribution that depends only on the
>
> $\underline{y}[1.\underline{P}, 2.\underline{P}]$ values and the $(\underline{S}, \underline{T})$ distribution. (3)

Fourth,

> Under $A_1$ and $A_2$,

$$E(D_1|\underline{S}=\underline{s},\underline{T}=\underline{t}) = E(D_1) = \mu_1 - \mu_2.$$

Under $B_1$ and $B_2$,

$$E(D_3|\underline{Y}=\underline{y},\underline{S}=\underline{s}) = E(D_3|\underline{S}=\underline{s}) = \overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]. \tag{4}$$

Under $C_1$ and $C_2$,

$$E(D_{23}|\underline{Y}=\underline{y}) = E(D_{23}) = \overline{y}[1.\underline{P}] - \overline{y}[2.\underline{P}].$$

Here, $\mu_t = E(\underline{Y}[t.i])$ is the mean of the assumed common distribution $F_t$. The last two expectation results follow because $D_3$ and $D_{23}$ are Horvitz–Thompson probability-weighted estimators (Horvitz and Thompson, 1952; Särndal et al., 1992, page 43). These expectation results highlight the usefulness of basing tests of "no treatment effects" on these $D$ statistics, at least when the treatment effect is measured in terms of differences in means. These results also highlight the usefulness of random sampling and treatment randomization.

### 6.2 Process-Based Tests

With the process-based approach, we condition on the selection (only $\underline{Y}$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]|(\underline{S}=\underline{s},\underline{T}=\underline{t}) \sim \underline{Y}[\underline{t}.\underline{s}]|(\underline{S}=\underline{s},\underline{T}=\underline{t})$$

to carry out inferences about the distribution of $\underline{Y}$. Among other assumptions, the validity of the process-based tests described below generally require that assumptions $A_1$: $(\underline{S},\underline{T}) \perp\!\!\!\perp \underline{Y}$; $A_2$: $\underline{Y}[t.i] \sim F_t, i = 1,\ldots,N, t = 1,2$; and $A_3$: $\underline{Y}[1.i, 2.i], i = 1,\ldots,N$ *are independent* hold. As noted in Section 5.1, these assumptions are often untenable in practice,[4] so the reader is reminded to apply these tests with caution.

6.2.1 *Permutation test*. Consider the no-treatment-effect hypothesis

$$H_0^{DUP} : \underline{Y}[1.i] \sim \underline{Y}[2.i], \quad i = 1,\ldots,N.$$

Under $H_0 = (A_1, A_2, A_3, H_0^{DUP})$, we can state the null as $H_0^{DUP}$: $F_1 = F_2$ and base our test on

$$D_1|\big(\underline{Y}[\underline{t}.\underline{s}] \in \Pi\big(\underline{y}[\underline{t}.\underline{s}]\big)\big)$$

which has a known, computable distribution under $H_0$. Here $\Pi(\underline{x}) = \{\text{set of distinct permutations of } \underline{x}\}$. That this distribution is known under $H_0$ follows because in this case

$$\underline{Y}[\underline{t}.\underline{s}]|\big(\underline{Y}[\underline{t}.\underline{s}] \in \Pi\big(\underline{y}[\underline{t}.\underline{s}]\big)\big) \overset{H_0}{\sim} \text{uniform over points in } \Pi\big(\underline{y}[\underline{t}.\underline{s}]\big). \tag{5}$$

The computability follows because $D_1(\underline{x})$ can be computed for any $\underline{x} \in \Pi(\underline{y}[\underline{t}.\underline{s}])$.

In practice, we would report a one- or two-sided $p$-value. For example, letting $D_{1,\text{obs}} = D_1(\underline{y}[\underline{t}.\underline{s}])$ be the observed difference, a two-sided $p$-value can be defined as

$$\text{pval}(D_{1,\text{obs}}) = P_{H_0}\big(|D_1| \geq |D_{1,\text{obs}}||\underline{Y}[\underline{t}.\underline{s}] \in \Pi\big(\underline{y}[\underline{t}.\underline{s}]\big)\big).$$

The size of the test that rejects $H_0$ iff pval $\leq \alpha$ is less than or equal to $\alpha$. If we observe a $p$-value $\leq \alpha$ and we assume that $A_1$, $A_2$, and $A_3$ hold, then we have statistical evidence against $H_0^{DUP}$: $F_1 = F_2$, that is, evidence at the $\alpha$ level that $F_1 \neq F_2$.

Remark: At first glance, one might think that exchangeability of the $N$ pairs $\underline{Y}[1.i, 2.i]$ could replace $(A_2, A_3)$. Unfortunately, a stronger exchangeability assumption would be needed to guarantee the uniform permutation distribution of (5). Specifically, the assumption must lead to the exchangeability of the $n$ components of $\underline{Y}[\underline{t}.\underline{s}]$. Along these lines, we could consider a more restrictive no-treatment-effect hypothesis, for example, $H_0^{DUP*}$: all $2N$ components in $\underline{Y}[1.\underline{P}, 2.\underline{P}]$ are exchangeable. Then the permutation test would be valid under $H_0^* = (A_1, H_0^{DUP*})$. It is useful to note that $H_0^{DUP*}$ can be viewed as $H_0^{DUP}$ along with extra assumptions about the process distribution. In this sense, this strong exchangeability hypothesis is an example of the no-treatment-effect hypotheses considered herein.

6.2.2 *Wilcoxon rank sum test*. Consider the no-treatment-effect hypothesis

$$H_0^{DUP} : \underline{Y}[1.i] \sim \underline{Y}[2.i], \quad i = 1,\ldots,N.$$

Under $H_0 = (A_1, A_2, A_3, A_4, H_0^{DUP})$, where $A_4$ is the assumption that the c.d.f.s $F_t$ are continuous, we can state the null as $H_0^{DUP}$: $F_1 = F_2$ and base our test on

$$W \equiv W(\underline{R}) = \sum_{j=1}^{n} R_j \mathbf{1}(t_j = 1),$$

where $W|\big(\underline{R} \in \Pi(\underline{r})\big)$ has a known, computable distribution under $H_0$.

[4] The tests are often invalid because the selection is related to the process, the treatment-specific process variables are not identically distributed, and/or the process variables are not independent across units.

Here $R_j = \text{rank}(\underline{Y}[t_j.s_j])$ and $r_j = \text{rank}(\underline{y}[t_j.s_j])$, where the ranks are taken over the $n$ values in $\underline{Y}[\underline{t}.\underline{s}]$ and $\underline{y}[\underline{t}.\underline{s}]$, respectively. Again, the $\Pi(\underline{r})$ is the set of permutations of $\underline{r}$. That this distribution is known under $H_0$ follows because in this case

$$(6)\quad \underline{R}|\big(\underline{R} \in \Pi(\underline{r})\big) \overset{H_0}{\sim} \text{uniform over points in } \Pi(\underline{r}).$$

The computability follows because $\underline{R}(\underline{x})$ can be computed for any $\underline{x} \in \Pi(\underline{r})$.

In practice, we would report a one- or two-sided $p$-value. For convenience, let $W_{\text{obs}} = W(\underline{r})$ be the observed rank sum statistic. Then the following is a reasonable two-sided $p$-value (of course there are others):

$$\text{pval}(W_{\text{obs}}) = 2\min\big\{P_{H_0}\big(W \geq W_{\text{obs}}|\underline{R} \in \Pi(\underline{r})\big), P_{H_0}\big(W \leq W_{\text{obs}}|\underline{R} \in \Pi(\underline{r})\big)\big\}.$$

Assuming that $A_1$–$A_4$ hold, an observed $p$-value $\leq \alpha$ would give us statistical evidence against $H_0^{DUP}$: $F_1 = F_2$, that is, evidence at the $\alpha$ level that $F_1 \neq F_2$.

6.2.3 *Two-sample $t$ tests*. Consider the no-treatment-effect hypothesis

$$H_0^{EUP} : E\big(\underline{Y}[1.i]\big) \sim E\big(\underline{Y}[2.i]\big), \quad i = 1, \ldots, N.$$

Under $H_0 = (A_1, A_2, A_3, A_5, H_0^{EUP})$, where $A_5$ states that we are sampling from $N(\mu_t, \sigma_t^2)$ distributions, we can state the null as $H_0^{EUP}$: $\mu_1 = \mu_2$ and base our test on

$$T \equiv \frac{D_1}{SE(D_1)} \overset{H_0}{\sim} \text{approx}\, t(\nu),$$

where $\nu$ is Welch's formula for the approximate degrees of freedom and $t(\nu)$ is Student's $t$ distribution. The standard error has the familiar form

$$SE(D_1) = \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}},$$

where $\widehat{\sigma}_t^2$ is the sample variance of the $\{\underline{Y}[t.s_j] : t_j = t\}$. Because $D_1$ is simply the difference between the two unweighted sample averages, this statistic $T$ is identical to Welch's (1938) version of the two-sample $t$ statistic.

An approximate two-sided $p$-value can be computed as

$$P_{H_0}(|T| \geq |T_{\text{obs}}|) \approx P\big(|t(\nu)| \geq |T_{\text{obs}}|\big) \equiv \text{apval}(T_{\text{obs}}).$$

Assuming that $A_1$–$A_3$ and $A_5$ hold, an approximate $p$-value $\leq \alpha$ gives statistical evidence against $H_0^{EUP}$: $\mu_1 = \mu_2$, that is, evidence at the approximate $\alpha$ level that $\mu_1 \neq \mu_2$.

Under $H_0 = (A_1, A_2, A_3, A_6, H_0^{EUP})$, where $A_6$ states that we are sampling from $N(\mu_t, \sigma^2)$ distributions, we can state the null as $H_0^{EUP}$: $\mu_1 = \mu_2$ and base our test on

$$T_p \equiv \frac{D_1}{SE_p(D_1)} \overset{H_0}{\sim} t(\nu), \qquad \nu = n_1 + n_2 - 2.$$

The standard error has the familiar form

$$SE_p(D_1) = \sqrt{\frac{\widehat{\sigma}^2}{n_1} + \frac{\widehat{\sigma}^2}{n_2}},$$

where $\widehat{\sigma}^2$ is the pooled estimate $((n_1 - 1)\widehat{\sigma}_1^2 + (n_2 - 1)\widehat{\sigma}_2^2)/(n_1 + n_2 - 2)$. The statistic $T_p$ is the standard two-sample pooled $t$ statistic.

An exact two-sided $p$-value can be computed as

$$P_{H_0}\big(|T_p| \geq |T_{p,\text{obs}}|\big) = P\big(|t(\nu)| \geq |T_{p,\text{obs}}|\big) \equiv \text{pval}(T_{p,\text{obs}}).$$

Assuming that $A_1$–$A_3$ and $A_6$ hold, an exact $p$-value $\leq \alpha$ gives statistical evidence against $H_0^{EUP}$: $\mu_1 = \mu_2$, that is, evidence at the approximate $\alpha$ level that $\mu_1 \neq \mu_2$.

Under the less restrictive assumption, $H_0 = (A_1, A_2, A_3, A_7, H_0^{EUP})$, where $A_7$ states that we are sampling from any distributions $F_t$ with mean $\mu_t$ and variance $\sigma_t^2$, we can still use $T$ to test $H_0^{EUP}$: $\mu_1 = \mu_2$, but the actual size of the tests based on the $p$-value, which uses the $t$ approximation, may be far from the nominal $\alpha$. In practice, the approximation is usually reasonable when $n$ is large enough to compensate for any asymmetry in the underlying $F_t$ distributions.

### 6.3 Randomization-Based Tests

With the randomization-based approach, we condition on both the process and the sample (only $\underline{T}$ is random), and use

$$\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]|(\underline{Y} = \underline{y}, \underline{S} = \underline{s}) \sim \underline{y}[\underline{T}.\underline{s}]|(\underline{Y} = \underline{y}, \underline{S} = \underline{s})$$

to carry out inferences about $\underline{y}[1.\underline{s}, 2.\underline{s}]$. The randomization-based test procedures outlined below are valid provided the assumptions $B_1$ and $B_2$ of Section 5.2 hold. There it was pointed out that these two assumptions can be made tenable when the treatment assignment is carried out by mechanical randomization.

6.3.1 *Fisher randomization test*. Although R. A. Fisher never explicitly used potential variables, several authors, including Welch (1937), Rubin (1990, 2005), and Cox (2009), have suggested that he tacitly used the no-unit-specific-effects (or sharp null) hypothesis in this randomized comparative experiment setting. That is, it has been suggested that to Fisher the no-treatment-effect hypothesis had the form $H_0^{RUs}$: $\underline{y}[1.s_j] = \underline{y}[2.s_j]$, $j = 1, \ldots, n$, or, in simpler notation,

$$H_0^{RUs} : \underline{y}[1.\underline{s}] = \underline{y}[2.\underline{s}].$$

When $H_0 = (B_1, B_2, H_0^{RUs})$ holds, we have by (4) that $E(D_3|\underline{S} = \underline{s}) \overset{H_0}{=} 0$ and we can base a test of $H_0$ on

$$D_3|(\underline{S} = \underline{s})$$

which has a known, computable distribution under $H_0$.

This null distribution is known because $D_3$ has form $D_3 = D_3(\underline{T})$ and assumption $B_2$ tells us that the distribution of $\underline{T}|(\underline{S} = \underline{s})$ is known. It is computable because under $B_1$, the distribution of $D_3|(\underline{S} = \underline{s})$ depends only on $\underline{y}[1.\underline{s}, 2.\underline{s}]$ and under $H_0^{RUs}$ the observed data $\underline{y}[\underline{t}.\underline{s}]$ determines the collection $\underline{y}[1.\underline{s}, 2.\underline{s}]$.

An exact two-sided $p$-value can be computed as

$$\begin{aligned}&\text{pval}(D_{3,\text{obs}})\\ &\quad = P_{H_0}(|D_3| \geq |D_{3,\text{obs}}||\underline{S} = \underline{s})\\ &\quad = P_{H_0}\big(\underline{T} \in \{\underline{x} : |D_3(\underline{x})| \geq |D_{3,\text{obs}}|\}|\underline{S} = \underline{s}\big).\end{aligned}$$

If we assume that $B_1$ and $B_2$ hold, an exact $p$-value $\leq \alpha$ gives statistical evidence against $H_0^{RUs}$: $\underline{y}[1.\underline{s}] = \underline{y}[2.\underline{s}]$, that is, evidence at the $\alpha$ level that $\underline{y}[1.s_j] \neq \underline{y}[2.s_j]$ for at least one subject $s_j$. This test is called a Fisher randomization test because it is based on the randomization approach and it was described by Fisher (1935).

This Fisher randomization test based on $D_3$ is tailored to detect differences between $\overline{y}[1.\underline{s}]$ and $\overline{y}[2.\underline{s}]$. To detect other differences, such as scale differences between the $\underline{y}[1.\underline{s}]$ and $\underline{y}[2.\underline{s}]$, an alternative to $D_3$ should (and can easily) be used.

Attractive features of this Fisher randomization test include the following: it has size guaranteed to be no larger than $\alpha$; it is valid when the sampling depends on the process ($\underline{S} \not\perp\!\!\!\perp \underline{Y}$); it does not require a model for the process variables $\underline{Y}$; and it does not require an estimate of the variance, $\text{var}(D_3|\underline{S} = \underline{s})$.

*Randomization vs. Permutation P-values*: It is clear that this Fisher randomization test is conceptually very different from the process-based permutation test. Indeed, as a rule, the randomization $p$-value based on $D_3$ is numerically different than the permutation $p$-value based on $D_1$. In fact, even if we had based both $p$-values on the same statistic $D_1$, the $p$-values would generally be different. There is an exception to this rule. Consider the special case uniform randomization distribution,

$$P(\underline{T} = \underline{x}|\underline{S} = \underline{s}) = \frac{\mathbf{1}(\underline{x} \in \Pi(\underline{t}))}{\binom{n}{n_1}}, \tag{7}$$

where $n_1$ is the number of 1's in $\underline{t}$ and $\Pi(\underline{t})$ is the set of all rearrangements of $n_1$ 1's and $n_2 = n - n_1$ 2's. This is the randomization used in the special case two-treatment *completely randomized design* (e.g., Cox, 1958b, pages 71–72; Kempthorne, 1977, Section 8). In this case, $D_1$ and $D_3$ are numerically identical, and the randomization and permutation $p$-values are numerically identical. It is this identity that often leads practitioners to incorrectly conclude that the process-based permutation test is identical to the randomization test. See Ernst (2004) for an interesting discussion.

6.3.2 *Neyman randomization test*. Compared to the view attributed to Fisher, Neyman was more interested in detecting nonzero treatment effects of the aggregate variety, especially $\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$. He apparently found it less practically useful to detect unit-specific effects if the average effect was 0. For this reason, Neyman used the no-average-effect hypothesis (cf. Welch, 1937). That is, Neyman viewed the no-treatment-effect hypothesis as

$$H_0^{RAs} : \overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}].$$

Because $H_0^{RAs} \supset H_0^{RUs}$, Neyman's approach focused on a narrower set of alternatives than Fisher, thereby opening up the possibility of finding a test with higher power than the Fisher randomization test, at least for alternatives of practical (in Neyman's view) interest.

When $H_0 = (B_1, B_2, H_0^{RAs})$ holds, we have by (4) that $E(D_3|\underline{S} = \underline{s}) \overset{H_0}{=} 0$ and we can consider basing a test of $H_0$ on

$$D_3|(\underline{S} = \underline{s})$$

which has a known, but *noncomputable*,

distribution under $H_0$.

This null distribution is known because $D_3 = D_3(\underline{T})$ and $\underline{T}|(\underline{S} = \underline{s})$ has a known distribution. It, however, is not computable because it depends on $\underline{y}[1.\underline{s}, 2.\underline{s}]$, which is not determined by the observed data $\underline{y}[\underline{t}.\underline{s}]$ under the no-average-effect hypothesis $H_0^{RAs}$. In contrast, recall that under the more restrictive unit-specific

(or sharp) null $H_0^{RUs}$, the observed data did determine $\underline{y}[1.\underline{s}, 2.\underline{s}]$.

Neyman was clearly aware of this noncomputability issue and instead invoked a central limit theorem and used

$$Z_3 \equiv Z_3(\underline{T}) \equiv \frac{D_3}{SE(D_3|\underline{S}=\underline{s})}$$

$$\text{where } Z_3|(\underline{S}=\underline{s}) \overset{H_0}{\sim} \text{approx } N(0,1).$$

Here, $SE$ is a standard error, which is an estimator of the standard deviation, $\text{sd}(D_3|\underline{S}=\underline{s})$. The standard deviation can be computed using sampling theory as described in Särndal et al. (1992). However, finding a reasonable estimator $SE$ of this standard deviation is more difficult because of the 0 second-order inclusion probabilities. Toward this end, Neyman (1923) derived a reasonable estimator of a tight upper bound for the variance under simplifying assumptions on the inclusion probabilities (Rubin, 1990, Gadbury, 2001; see Copas, 1973, for a related result). It is useful to note that the variance attains this upper bound when unit-treatment additivity holds, that is, $\underline{y}[1.s_j] = \underline{y}[2.s_j] + \text{constant}$, $j = 1, \ldots, n$. In this paper, we use the Neyman estimator of variance. The square root of this estimator is $SE(D_3|\underline{S}=\underline{s})$.

REMARK. There is a related approximate normality result when $H_0^{RAs}$ does not hold. Under $(B_1, B_2)$, we noted in (4) that $E(D_3|\underline{S}=\underline{s}) = \overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$ and we have that $\text{sd}(D_3|\underline{S}=\underline{s})$ is approximated by $SE(D_3|\underline{S}=\underline{s})$. By the central limit theorem and continuous mapping results, we have

$$\frac{D_3 - (\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}])}{SE(D_3|\underline{S}=\underline{s})}|(\underline{S}=\underline{s}) \sim \text{approx } N(0,1).$$

This result is useful for testing other hypotheses and for computing confidence intervals.

The Normal approximation for $Z_3$ generally improves as the number of support points in $\underline{T}|(\underline{S}=\underline{s})$ increases. However, when the differences $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ are highly variable, the unit variance in the approximation can be a substantial overestimate (see Gadbury, 2001), and when $\underline{y}[1.s_j] - \underline{y}[2.s_j] = \text{constant}$, the unit variance can be a slight underestimate when the sample sizes are small (based on observations from the simulation study carried out for this paper).

An approximate two-sided $p$-value can be computed as

$$P_{H_0}(|Z_3| \geq |Z_{3,\text{obs}}||\underline{S}=\underline{s}) \approx P(|N(0,1)| \geq |Z_{3,\text{obs}}|)$$
$$\equiv \text{apval}(Z_{3,\text{obs}}).$$

If $B_1$ and $B_2$ hold, an approximate $p$-value $\leq \alpha$ gives statistical evidence against $H_0^{RAs}$: $\overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}]$, that is, evidence at the approximate $\alpha$ level that $\overline{y}[1.\underline{s}] \neq \overline{y}[2.\underline{s}]$. This test is called a Neyman randomization test because it is based on the randomization approach and ideas in Neyman (1923).

Unlike the Fisher randomization test of $H_0^{RUs}$, the size of the Neyman test of $H_0^{RAs}$ is not guaranteed to be less than or equal to $\alpha$; it is only approximately size $\alpha$. For smaller $n_1$ and $n_2$ and when the more restrictive hypothesis $H_0^{RUs}$ holds, the Neyman randomization test tends to be anti-conservative, with size a bit larger than the nominal $\alpha$. This follows because the Neyman estimator of the variance tends to slightly underestimate the true variance in this case. For moderate $n_1$ and $n_2$ the approximation is usually reasonable provided $D_3(\underline{T})$ has enough support points with respect to the $\underline{T}|(\underline{S}=\underline{s})$ distribution. We empirically explore this approximation below.

### 6.4 Selection-Based Tests

With the selection-based approach, we condition on the process values [only $(\underline{S}, \underline{T})$ is random] and use

$$\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]|(\underline{Y}=\underline{y}) \sim \underline{y}[\underline{T}.\underline{S}]|(\underline{Y}=\underline{y})$$

to carry out inferences about $\underline{y}[1.\underline{P}, 2.\underline{P}]$. The selection-based test procedures outlined below are valid provided the assumptions $C_1$ and $C_2$ of Section 5.3 hold. There it was pointed out that these two assumptions are often untenable, so the following test procedures must be applied with caution.

6.4.1 *Fisher selection test.* The no-unit-specific-treatment-effect (or sharp null) hypothesis in this selection-based setting has the form $H_0^{RUP}$: $\underline{y}[1.i] = \underline{y}[2.i]$, $i = 1, \ldots, N$, or, more simply,

$$H_0^{RUP} : \underline{y}[1.\underline{P}] = \underline{y}[2.\underline{P}].$$

When $H_0 = (C_1, C_2, H_0^{RUP})$ holds, we have by (4) that $E(D_{23}) \overset{H_0}{=} 0$ and we can consider basing a test of $H_0$ on

$$D_{23}|(\underline{S}=\underline{s})$$

which has a known, but *noncomputable*

distribution under $H_0$.

This null distribution is known because $D_{23}$ has form $D_{23} = D_{23}(\underline{S}, \underline{T})$ and assumption $C_2$ tells us that the distribution of $(\underline{S}, \underline{T})$ is known. It is, however, not computable because it depends on $\underline{y}[1.\underline{P}, 2.\underline{P}]$, which is

not determined by the observed data $\underline{y}[\underline{t}.\underline{s}]$ under the hypothesis $H_0^{RUP}$. To see this, note that for $\underline{s}' \neq \underline{s}$, there is an $s'_j$ such that both $\underline{y}[1.s'_j]$ and $\underline{y}[2.s'_j]$ are unobserved and hence not computable even under $H_0^{RUP}$.

It follows that an exact Fisher selection test is *not* available in this selection-based setting. We could condition on the sample and be content using the Fisher randomization test to draw inferences about $\underline{y}[1.\underline{s}, 2.\underline{s}]$ rather than $\underline{y}[1.\underline{P}, 2.\underline{P}]$. Alternatively, we could use the approximate selection-based test described in the next subsection.

6.4.2 *Neyman selection test.* In analogy to the randomization setting, Neyman likely would consider the no-average-effect hypothesis:

$$H_0^{RAP} : \overline{y}[1.\underline{P}] = \overline{y}[2.\underline{P}].$$

When $H_0 = (C_1, C_2, H_0^{RAP})$ holds, we have by (4) that $E(D_{23}) \overset{H_0}{=} 0$ and, analogous to the randomization setting, we can base a test of $H_0$ on

$$Z_{23} \equiv Z_{23}(\underline{S}, \underline{T}) = \frac{D_{23}}{SE(D_{23})}$$

$$\text{where } Z_{23} \overset{H_0}{\sim} \text{approx } N(0,1).$$

Just as with $\text{sd}(D_3|\underline{S} = \underline{s})$ in the randomization approach, the standard deviation $\text{sd}(D_{23})$ can be computed and estimated using sampling theory. The estimation, however, is subject to the same problems as in the randomization approach because of the nonmeasurability of probability sample $\underline{T}.\underline{S}$. Suffice it to say that a reasonable Neyman estimator $SE(D_{23})$, analogous to the one in the randomization setting, exists.

The approximate Normality result follows just as in the randomization setting. Specifically, under $C_1$ and $C_2$, and using the same arguments as in the randomization approach, we have that quite generally

$$\frac{D_{23} - (\overline{y}[1.\underline{P}] - \overline{y}[2.\underline{P}])}{SE(D_{23})} \sim \text{approx } N(0,1).$$

The approximation generally improves as the number of support points in $\underline{T}.\underline{S}$ increases. However, when the differences $\underline{y}[1.i] - \underline{y}[2.i]$ are highly variable, the unit variance in the approximation can be a substantial overestimate (see Gadbury, 2001).

An approximate two-sided $p$-value can be computed as

$$P_{H_0}(|Z_{23}| \geq |Z_{23,\text{obs}}|) \approx P(|N(0,1)| \geq |Z_{23,\text{obs}}|) \equiv \text{apval}(Z_{23,\text{obs}}).$$

If $C_1$ and $C_2$ hold, an approximate $p$-value $\leq \alpha$ gives statistical evidence against $H_0^{RAP}$: $\overline{y}[1.\underline{P}] = \overline{y}[2.\underline{P}]$, that is, evidence at the approximate $\alpha$ level that $\overline{y}[1.\underline{P}] \neq \overline{y}[2.\underline{P}]$. This test is called a Neyman selection test because it is based on the selection approach and ideas in Neyman (1923).

Just as in the randomization setting, the size of the Neyman test of $H_0^{RAP}$ is not guaranteed to be less than or equal to $\alpha$; it is only approximately size $\alpha$. Remarks regarding the approximation in this selection setting are analogous to those given at the end of Section 6.3.2, in the randomization setting.

## 7. EMPIRICAL INVESTIGATIONS

### 7.1 Cell Phone Use Example (Revisited)

The process variable $\underline{Y}[t.i]$ is defined as the reaction time for the $i$th unit in population $\underline{P}$ when exposed to treatment $t$. Inference about the process $\underline{Y}$ distribution will be difficult to describe because the sample of 64 students was not taken from any well-defined population $\underline{P}$. For any substantively interesting population, for example, $\underline{P}$ = licensed drivers in Utah, the assumption that $\underline{S} \perp\!\!\!\perp \underline{Y}$ is untenable given the haphazard nature of the sample selection. The untenability of $\underline{S} \perp\!\!\!\perp \underline{Y}$ also implies that it will be difficult to carry out inferences about the population values $\underline{y}[1.\underline{P}, 2.\underline{P}]$ for any substantively interesting population $\underline{P}$. For these reasons, it makes sense to focus on inferences about the 128 potential values in $\underline{y}[1.\underline{s}, 2.\underline{s}]$. That is, it is arguably better to use randomization-based inference for this example.

We assume that the randomization was carried out mechanically so that $\underline{T} \perp\!\!\!\perp \underline{Y}|\underline{S}$ and we assume that the distribution of $\underline{T}|(\underline{S} = \underline{s})$ is uniform in the sense of (7); that is, conditions $B_1$ and $B_2$ of Section 6.3 are assumed to hold. We will use the Fisher randomization test to test the no-treatment-effect hypothesis $H_0^{RUs}$: $\underline{y}[1.s_j] = \underline{y}[2.s_j]$, $j = 1, \ldots, 64$ and the Neyman randomization test to test the no-treatment-effect hypothesis $H_0^{RAs}$: $\overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}]$.

For these data, the observed randomization statistics are

$$D_{3,\text{obs}} = 51.59, \quad Z_{3,\text{obs}} = \frac{51.59}{19.30} = 2.67,$$

$$\text{pval}(D_{3,\text{obs}}) = 0.0074 \quad \text{and}$$

$$\text{apval}(Z_{3,\text{obs}}) = 0.0075.$$

Because the Fisher randomization $p$-value $\text{pval}(D_{3,\text{obs}}) = 0.0074$ is small, we have sufficient evidence to reject $H_0^{RUs}$; there is statistical evidence

that $\underline{y}[1.s_j] \neq \underline{y}[2.s_j]$ for at least one subject in the sample of 64. Because the Neyman randomization $p$-value $\text{apval}(Z_{3,\text{obs}}) = 0.0075$ is small, we have sufficient evidence to reject $H_0^{RAs}$; there is statistical evidence that $\overline{y}[1.\underline{s}] \neq \overline{y}[2.\underline{s}]$. In fact, because $D_{3,\text{obs}} = 51.59$ is a Horvitz–Thompson unbiased estimate of $\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$, the Neyman test gives statistical evidence that the reaction time values are higher on average when cell phones are used, at least for this sample of 64. In other words, there is statistical evidence of a treatment effect.

For completeness and for comparison purposes, we also give the values of the other commonly used $p$-values, viz., permutation, Wilcoxon, Welch's approximate $t$, and the pooled $t$:

$$\text{pval}(D_{1,\text{obs}}) = 0.0074, \quad \text{pval}(W_{\text{obs}}) = 0.0184,$$
$$\text{apval}(T_{\text{obs}}) = 0.0110 \quad \text{and} \quad \text{pval}(T_{p,\text{obs}}) = 0.0107.$$

Strictly speaking, these are only applicable for process-based inference, so they are of questionable utility for this example. As noted above, because the randomization distribution is uniform, the permutation $p$-value $\text{pval}(D_{1,\text{obs}})$ is numerically (but not conceptually!) identical to the Fisher randomization $p$-value $\text{pval}(D_{3,\text{obs}})$.

All the computations were carried out in R. The author has written code to compute the Neyman randomization $p$-value. The Fisher randomization and permutation $p$-values were approximated using Monte-Carlo estimation (here we used $10^6$ simulations) as carried out in `twot.permutation {DAAG}`. The Wilcoxon $p$-value was computed using `wilcox.test {stats}`. Note that when there are ties, as there are in this example, `wilcox.test` only reports approximate $p$-values.

### 7.2 A Simulation Study

This section empirically compares the operating characteristics of the different tests considered in this paper, under a variety of scenarios. All computations were carried out in R, with $p$-values computed as described at the end of the previous subsection. The simulated data are generated according to models of the form

$$\begin{aligned}
&\underline{y}[1.i] \leftarrow \underline{Y}[1.i] \text{ IID} \sim [\text{scenario}],\\
&\underline{y}[2.i] \leftarrow \underline{Y}[2.i] \sim [\text{scenario}], \qquad i = 1, \ldots, N,\\
&\underline{s} \leftarrow \underline{S}|(\underline{Y} = \underline{y}) \sim P\big(\underline{S} = (1, \ldots, n)|\underline{Y} = \underline{y}\big) = 1,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{where } n = N,
\end{aligned} \tag{8}$$

$$\begin{aligned}
\underline{t} \leftarrow \underline{T}|(\underline{Y} = \underline{y}, \underline{S} = \underline{s}) &\sim P\big(\underline{T} = \underline{t}'|\underline{Y} = \underline{y}, \underline{S} = \underline{s}\big)\\
&= \frac{n_1!n_2!}{n!}\mathbf{1}(\underline{t}' \in \mathcal{T}),
\end{aligned}$$

where $n = n_1 + n_2$ and $\mathcal{T}$ is the set of all possible rearrangements of $n_1$ 1's and $n_2$ 2's. Looking back at the process-based assumptions of Section 5.1, we see that $A_1$ holds, but none of $A_2$–$A_7$ is guaranteed to hold. Both the randomization-based assumptions $B_1$ and $B_2$ of Section 5.2 hold, as do both the selection-based assumptions $C_1$ and $C_2$ of Section 5.3. A more extensive simulation would also investigate scenarios where more of the assumptions do not hold.

For data-generation models of the form (8), we have that (i) the randomization- and selection-based approaches are identical because the sample $\underline{S}$ is taken to be equal to the population $\underline{P}$ with probability one; and (ii) the permutation and Fisher randomization $p$-values are numerically (not conceptually!) identical because the randomization distribution (the distribution of $\underline{T}$) is uniform over the set of all possible treatment assignments.

Although the permutation-, Wilcoxon-, and $t$-tests are process-based approaches, we will estimate their operating characteristics for both the process and randomization (here, randomization = selection) distributions. Similarly, the Fisher and Neyman randomization tests are randomization-based approaches, but we report their operating characteristics for both the process and the randomization distributions. In the tables below, the rows labeled "Randomization" give Monte-Carlo estimates of the power of the tests over the distribution $\underline{T}|(\underline{Y} = \underline{y}, \underline{S} = \underline{s})$. The rows labeled "Process" give Monte-Carlo estimates of the power of the tests over the distribution $\underline{Y}|(\underline{S} = \underline{s}, \underline{T} = \underline{t})$. In all cases, the nominal size is set at $\alpha = 0.05$.

The simulation results in Tables 3–6 give us a glimpse at the operating characteristics of the tests for a variety of scenarios, labeled "Sc. #." The following summary focuses on comparisons between the Fisher and Neyman randomization tests, but the table entries afford broader comparisons.

For small $n_1, n_2$, when $\underline{y}[1.s_j] - \underline{y}[2.s_j] = \text{constant}$, the Neyman randomization test tends to be just a bit anti-conservative for testing $H_0^{RAs}$; that is, the actual size appears to be a little larger than the nominal size (see scenarios 1, 2, and 6 of Table 3). This anti-conservativeness presumably stems from the fact that the Neyman estimator of the variance, $\text{var}(D_3|\underline{S} = \underline{s})$, tends to be slightly biased on the low side when

TABLE 3
*Monte-Carlo estimates of size when $n_1 = n_2 = 10$, nominal size = 5%*

| $n_1 = n_2 = 10$ | Permutation[a] | Wilcoxon | $t$(Welch) | $t$(Pooled) | Fisher[a] | Neyman | |
|---|---|---|---|---|---|---|---|
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$ | | | | | | Sc. 1 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 20$ | | | | | | |
| Randomization | 4.6 | 3.6 | 4.7 | 4.7 | 4.6 | 6.5 | |
| Process | 4.3 | 3.4 | 4.2 | 4.3 | 4.3 | 6.9 | |
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$) | | | | | | Sc. 2 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 20$ | | | | | | |
| Randomization | 5.0 | 4.9 | 4.1 | 4.6 | 5.0 | 7.4 | |
| Process | 4.0 | 4.1 | 3.2 | 3.5 | 4.0 | 7.7 | |
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim 0.9U(0, 20) + 0.1U(200, 201)$, "mixture of uniforms" | | | | | | Sc. 3 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 20$ | | | | | | |
| Randomization[b] | 4.6 | 3.9 | 0.0 | 0.0 | 4.6 | 0.0 | |
| Process | 3.8 | 3.5 | 1.1 | 1.8 | 3.8 | 11.2 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$ | | | | | | Sc. 4 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + E_i - \overline{E}$, $E_i$ IID $\sim N(0, 3^2), i = 1, \ldots, 20$ | | | | | | |
| Randomization | 1.5 | 1.9 | 1.7 | 1.8 | 1.5 | 3.3 | |
| Process | 2.7 | 2.0 | 2.5 | 2.6 | 2.7 | 4.2 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$) | | | | | | Sc. 5 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 2\underline{Y}[1.i] - \overline{Y}[1.\underline{P}], i = 1, \ldots, 20$ | | | | | | |
| Randomization | 4.8 | 6.8 | 4.3 | 4.4 | 4.8 | 7.6 | |
| Process | 4.0 | 7.4 | 3.6 | 3.7 | 4.0 | 7.4 | |
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID$\sim$ bin(1, 0.28) | | | | | | Sc. 6 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 20$ | | | | | | |
| Randomization[c] | 0.0 | NA | 9.1 | 9.1 | 0.0 | 9.1 | |
| Process | 2.1 | NA | 4.5 | 4.5 | 2.1 | 11.3 | |
| $H_0^{DUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID$\sim$[d] bin(1, 0.28) | | | | | | Sc. 7 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i]$ IID$\sim$[d] bin(1, 0.28), corr($\underline{Y}[1.i], \underline{Y}[2.i]) = 0.37, i = 1, \ldots, 20$ | | | | | | |
| Randomization[e] | 0.4 | NA[f] | 1.6 | 1.6 | 0.4 | 4.6 | |
| Process | 0.2 | NA | 1.0 | 1.0 | 0.2 | 3.7 | |

Table entries give the rejection rates (as a percent) for the 1000 simulations.

All indented hypotheses are also true; see Section 4.2. For example, in row 1, $H_0^{UP}$ is true. It follows that all the other hypotheses in Section 4.2 are also true.

[a]For this simulation, the permutation and Fisher randomization test results are numerically identical.

[b]The fixed $\underline{y}$ includes one large observation from the $U(200, 201)$ distribution.

[c]The fixed $\underline{y}[1.\underline{P}] = 0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,1\,0 = \underline{y}[2.\underline{P}]$.

[d]This is an approximation because the $\underline{Y}$ values are adjusted to satisfy $H_0^{RAs}$.

[e]The fixed $\underline{y}[1.\underline{P}] = 1\,0\,0\,1\,0\,1\,0\,1\,0\,0\,1\,0\,0\,1\,0\,0\,0\,0\,0\,0$, $\underline{y}[2.\underline{P}] = 0\,0\,0\,0\,1\,1\,0\,1\,0\,0\,1\,0\,0\,1\,1\,0\,0\,0\,0\,0$.

[f]Because of the many ties in the binomial case, the Wilcoxon test as described herein is not applicable.

$\underline{y}[1.s_j] - \underline{y}[2.s_j] =$ constant. For larger $n_1, n_2$, this anti-conservativeness disappears (scenarios 1, 2, and 6 of Table 5).

When the differences $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ are highly variable, the Neyman randomization test tends to be a bit conservative for testing $H_0^{RAs}$, although not as conservative as the Fisher randomization test (scenarios 4 and 7 in Tables 3 and 5). This conservativeness presumably stems from the fact that the Neyman estimator of the variance, var$(D_3|\underline{S} = \underline{s})$, tends to be biased on the high side when $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ are highly variable (see Gadbury, 2001).

TABLE 4
*Monte-Carlo estimates of power when $n_1 = n_2 = 10$, nominal size = 5%*

| $n_1 = n_2 = 10$ | Permutation[a] | Wilcoxon | $t$(Welch) | $t$(Pooled) | Fisher[a] | Neyman | |
|---|---|---|---|---|---|---|---|
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$ | | | | | | Sc. 1 |
| false | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 2, i = 1, \ldots, 20$ | | | | | | |
| Randomization | 52.7 | 49.3 | 51.3 | 52.5 | 52.7 | 59.9 | |
| Process | 55.9 | 51.6 | 55.5 | 56.1 | 55.9 | 62.7 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$ | | | | | | Sc. 2 |
| false | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 2 + E_i - \overline{E}$, $E_i$ IID $\sim N(0, 3^2), i = 1, \ldots, 20$ | | | | | | |
| Randomization | 26.2 | 23.6 | 24.1 | 25.7 | 26.2 | 35.6 | |
| Process | 28.6 | 23.8 | 26.1 | 27.2 | 28.6 | 36.6 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$ | | | | | | Sc. 3 |
| false | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.2\underline{Y}[1.i], i = 1, \ldots, 20$ | | | | | | |
| Randomization | 34.7 | 27.1 | 34.8 | 35.3 | 34.7 | 43.0 | |
| Process | 48.4 | 43.0 | 47.5 | 48.4 | 48.4 | 57.2 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$) | | | | | | Sc. 4 |
| false | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 2\underline{Y}[1.i], i = 1, \ldots, 20$ | | | | | | |
| Randomization | 19.2 | 12.9 | 16.1 | 18.7 | 19.2 | 28.6 | |
| Process | 30.2 | 23.7 | 23.4 | 26.0 | 30.2 | 38.5 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$) | | | | | | Sc. 5 |
| false | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 3\underline{Y}[1.i] + E_i$, $E_i$ IID $\sim N(0, 5^2), i = 1, \ldots, 20$ | | | | | | |
| Randomization | 45.7 | 28.6 | 40.2 | 45.3 | 45.7 | 65.5 | |
| Process | 49.2 | 38.1 | 39.8 | 44.4 | 49.2 | 63.5 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID$\sim$ bin$(1, 0.28)$ | | | | | | Sc. 6 |
| false | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i]$ IID$\sim$ bin$(1, 0.71)$, corr$(\underline{Y}[1.i], \underline{Y}[2.i]) = 0.29, i = 1, \ldots, 20$ | | | | | | |
| Randomization[b] | 18.9 | NA | 37.3 | 37.3 | 18.9 | 37.4 | |
| Process | 29.6 | NA | 48.0 | 48.0 | 29.6 | 50.3 | |

Table entries give the rejection rates (as a percent) for the 1000 simulations.
[a]For this simulation, the permutation and Fisher randomization test results are numerically identical.
[b]The fixed $\underline{y}[1.\underline{P}] = 0\,0\,0\,0\,0\,1\,0\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,1\,0$, $\underline{y}[2.\underline{P}] = 0\,1\,1\,1\,1\,1\,1\,0\,1\,0\,1\,1\,0\,0\,1\,1\,1\,0\,1\,1$.

For small $n_1, n_2$, the Normal approximation to the Neyman test statistic can be unreasonable when there are extreme outliers present (scenario 3 of Table 3). With larger $n_1, n_2$, the Normal approximations become more reasonable in the presence of extreme outliers (scenario 3 of Table 5).

In all of the simulation scenarios, the Neyman randomization test had higher power than the Fisher randomization test (see Tables 4 and 6), especially when $n_1, n_2$ are smaller (see Table 4). Of course, power comparisons are most useful when both tests have the same size. Because neither of these tests has size exactly equal to the nominal 0.05, these power comparisons should be considered carefully. In particular, in head-to-head comparisons, the Fisher test is at a disadvantage because its actual size is guaranteed to be no larger than 0.05; the Neyman test has size that is only approximately equal to, and can exceed, the nominal 0.05.

On the basis of this limited simulation study, we recommend that practitioners at least think seriously about using the Neyman randomization test as an alternative to the Fisher randomization test, especially when $n_1, n_2$ are moderate, say, at least 10, and when there are no extreme outliers.

## 8. DISCUSSION

This paper used concepts from the rich literatures on causal analysis and finite-population sampling theory to clear up some of the confusion that exists about tests of the no-treatment-effect hypothesis in the randomized comparative experiment setting. Our approach lends itself to explicit specifications of the can-

TABLE 5
*Monte-Carlo estimates of size when $n_1 = n_2 = 50$, nominal size = 5%*

| $n_1 = n_2 = 50$ | Permutation[a] | Wilcoxon | $t$(Welch) | $t$(Pooled) | Fisher[a] | Neyman | |
|---|---|---|---|---|---|---|---|
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$ | | | | | | Sc. 1 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 100$ | | | | | | |
| Randomization | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.4 | |
| Process | 4.7 | 4.8 | 4.8 | 4.8 | 4.7 | 5.5 | |
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$) | | | | | | Sc. 2 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 100$ | | | | | | |
| Randomization | 4.9 | 5.0 | 4.8 | 4.8 | 4.9 | 5.4 | |
| Process | 4.1 | 3.9 | 3.9 | 3.9 | 4.1 | 4.8 | |
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim 0.9U(0, 20) + 0.1U(200, 201)$, "mixture of uniforms" | | | | | | Sc. 3 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 100$ | | | | | | |
| Randomization[b] | 4.2 | 6.5 | 4.3 | 4.5 | 4.2 | 8.6 | |
| Process | 5.3 | 5.4 | 5.3 | 5.3 | 5.3 | 6.6 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$ | | | | | | Sc. 4 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + E_i - \overline{E}$, $E_i$ IID $\sim N(0, 3^2), i = 1, \ldots, 100$ | | | | | | |
| Randomization | 2.5 | 3.1 | 2.4 | 2.4 | 2.5 | 3.4 | |
| Process | 3.0 | 4.6 | 2.9 | 3.2 | 3.0 | 3.9 | |
| $H_0^{EUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$) | | | | | | Sc. 5 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 2\underline{Y}[1.i] - \overline{Y}[1, \underline{P}], i = 1, \ldots, 100$ | | | | | | |
| Randomization | 4.6 | 42.5 | 4.4 | 4.4 | 4.6 | 6.1 | |
| Process | 2.8 | 35.1 | 2.8 | 2.8 | 2.8 | 5.0 | |
| $H_0^{UP}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID$\sim$bin$(1, 0.28)$ | | | | | | Sc. 6 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i], i = 1, \ldots, 100$ | | | | | | |
| Randomization[c] | 2.2 | NA | 5.5 | 5.5 | 2.2 | 5.5 | |
| Process | 3.5 | NA | 5.0 | 5.0 | 3.5 | 5.9 | |
| $H_0^{DUP}, H_0^{RAs}$ | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID$\sim$[d] bin$(1, 0.28)$ | | | | | | Sc. 7 |
| true | $\underline{y}[2.i] \leftarrow \underline{Y}[2.i]$ IID$\sim$[d] bin$(1, 0.28)$, corr$(\underline{Y}[1.i], \underline{Y}[2.i]) = 0.37, i = 1, \ldots, 100$ | | | | | | |
| Randomization[e] | 0.5 | NA | 0.8 | 0.8 | 0.5 | 1.0 | |
| Process | 1.6 | NA | 2.8 | 2.8 | 1.6 | 3.5 | |

Table entries give the rejection rates (as a percent) for the 1000 simulations.
[a]For this simulation, the permutation and Fisher randomization test results are numerically identical.
[b]The fixed $\underline{y}$ includes 7 large observations from the $U(200, 201)$ distribution.
[c]The fixed $\underline{y}[1.\underline{P}] = \underline{y}[2.\underline{P}]$ with $\overline{y}[1.\underline{P}] = \overline{y}[2.\underline{P}] = 32/100$.
[d]This is an approximation because the $\underline{Y}$ values are adjusted to satisfy $H_0^{RAs}$.
[e]The fixed $\underline{y}$ is such that $\underline{y}[1.\underline{P}] \neq \underline{y}[2.\underline{P}]$, $\overline{y}[1.\underline{P}] = \overline{y}[2.\underline{P}] = 33/100$, and corr$(\underline{y}[1.\underline{P}], \underline{y}[2.\underline{P}]) = 0.186$.

didate no-treatment-effects hypotheses and targets of inference. We clearly distinguished between three main inference approaches: process-based, randomization-based, and selection-based. The commonly-used permutation test, Wilcoxon rank sum test, and two-sample $t$ tests are examples of process-based approaches. Examples of randomization-based approaches include the commonly-used Fisher randomization test and the less commonly-used Neyman randomization test. We also described a Neyman selection test. A small-scale empirical comparison of these different tests was carried out. On the basis of the simulation results, we recommend that practitioners consider using the Neyman randomization test in certain scenarios.

In our description of the process-based approach, we focused on testing hypotheses about the distribution of $\underline{Y}$. More generally, the process-based approach can be used to both estimate, or test hypotheses about, char-

TABLE 6
*Monte-Carlo estimates of power when $n_1 = n_2 = 50$, nominal size = 5%*

| $n_1 = n_2 = 50$ | Permutation[a] | Wilcoxon | $t$(Welch) | $t$(Pooled) | Fisher[a] | Neyman | |
|---|---|---|---|---|---|---|---|
| $H_0^{EUP}, H_0^{RAs}$ false | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$<br>$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 1, i = 1, \ldots, 100$ | | | | | | Sc. 1 |
| Randomization | 80.9 | 76.4 | 80.4 | 80.4 | 80.9 | 81.3 | |
| Process | 69.5 | 67.7 | 69.9 | 69.9 | 69.5 | 70.4 | |
| $H_0^{EUP}, H_0^{RAs}$ false | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$<br>$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 1 + E_i - \overline{E}$, $E_i$ IID $\sim N(0, 3^2), i = 1, \ldots, 100$ | | | | | | Sc. 2 |
| Randomization | 36.3 | 31.4 | 36.2 | 36.4 | 36.3 | 42.7 | |
| Process | 37.9 | 36.3 | 37.5 | 38.0 | 37.9 | 42.8 | |
| $H_0^{EUP}, H_0^{RAs}$ false | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim N(10, 2^2)$<br>$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.1\underline{Y}[1.i], i = 1, \ldots, 100$ | | | | | | Sc. 3 |
| Randomization | 70.5 | 68.6 | 71.0 | 71.1 | 70.5 | 72.1 | |
| Process | 66.6 | 63.9 | 65.7 | 65.7 | 66.6 | 67.4 | |
| $H_0^{EUP}, H_0^{RAs}$ false | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$)<br>$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.5\underline{Y}[1.i], i = 1, \ldots, 100$ | | | | | | Sc. 4 |
| Randomization | 46.6 | 39.0 | 46.2 | 46.4 | 46.6 | 49.5 | |
| Process | 49.2 | 40.6 | 48.0 | 48.2 | 49.2 | 51.8 | |
| $H_0^{EUP}, H_0^{RAs}$ false | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID $\sim$ Gamma(shape $= 1$, scale $= 5$)<br>$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.5\underline{Y}[1.i] + E_i$, $E_i$ IID $\sim N(0, 5^2), i = 1, \ldots, 100$ | | | | | | Sc. 5 |
| Randomization | 41.0 | 35.2 | 40.4 | 40.5 | 41.0 | 44.3 | |
| Process | 39.2 | 30.7 | 38.9 | 39.0 | 39.2 | 44.2 | |
| $H_0^{EUP}, H_0^{RAs}$ false | $\underline{y}[1.i] \leftarrow \underline{Y}[1.i]$ IID$\sim$ bin$(1, 0.28)$<br>$\underline{y}[2.i] \leftarrow \underline{Y}[2.i]$ IID$\sim$ bin$(1, 0.50)$, corr$(\underline{Y}[1.i], \underline{Y}[2.i]) = 0.36, i = 1, \ldots, 100$ | | | | | | Sc. 6 |
| Randomization[b] | 48.8 | NA | 58.8 | 58.8 | 48.8 | 60.3 | |
| Process | 51.1 | NA | 60.1 | 60.1 | 51.1 | 60.3 | |

Table entries give the rejection rates (as a percent) for the 1000 simulations.
[a]For this simulation, the permutation and Fisher randomization test results are numerically identical.
[b]The fixed $\underline{y}$ is such that $\underline{y}[1.\underline{P}] \neq \underline{y}[2.\underline{P}]$, $\overline{y}[1.\underline{P}] = 24/100$, $\overline{y}[2.\underline{P}] = 45/100$, and corr$(\underline{y}[1.\underline{P}], \underline{y}[2.\underline{P}]) = 0.386$.

acteristics of the distribution of $\underline{Y}$ *and* predict/estimate the unobserved values $\underline{y}[-\underline{t}.\underline{s}]$. Here, $\underline{y}[-A]$ is the collection of all $2N$ components of $\underline{y}$ excluding those with subscripts in the set $A$. A look back at the assumptions $A_1$–$A_7$ shows that we did not have to specify a model for the joint distribution of $\underline{Y}$ to carry out a test of no treatment effect. We only assumed independence across units and modeled the marginal distributions of $\underline{Y}[1.i]$ and $\underline{Y}[2.i]$. In contrast, the prediction of unobserved values generally requires a model for the joint distribution of $\underline{Y}$, equivalently, a model for $(\underline{Y}[\underline{t}.\underline{s}], \underline{Y}[-\underline{t}.\underline{s}])$, the "$(Y_{\text{obs}}, Y_{\text{mis}})$" of Rubin (e.g., 2005). Rubin advocates using a Bayesian approach to process-based prediction of $\underline{y}[-\underline{t}.\underline{s}]$.

This paper restricted attention to inferences about one population or sample, under two scenarios corresponding to two treatments. Owing to randomization, we were able to compare these two treatment scenarios; for example, see equation (4). Comparing two populations of distinct units is a qualitatively different inference problem. However, similar notation and model structures can be used to study this problem as well. Interestingly, in this two population setting, Fisher randomization tests, as described herein, are generally not applicable. In contrast, the other tests described in this paper, including the Neyman selection test, are applicable.

The notation and model structure introduced in this paper can be directly applied in more general settings where nonuniform or constrained randomization is used or where there are more than two treatments being compared; see, for example, the descriptions in

Sutter et al. (1963), Kempthorne (1977), and Bailey (1981). There are extensions in other directions. For example, rather than testing hypotheses, the ideas introduced in this paper show promise for confidence interval estimation. More work in this direction will be forthcoming.

In the binary response, comparative experiment setting, *Fisher's exact test* for $2 \times 2$ tables (see Agresti, 2002, page 91) is equivalent to the Fisher randomization test of $H_0^{RUs}$ when $\underline{T} \perp\!\!\!\perp \underline{Y}|\underline{S}$ and $\underline{T}|(\underline{S} = \underline{s})$ have a uniform distribution as in (7); recall that $H_0^{RUs}$ states that the binary response values satisfy $\underline{y}[1.s_j] = \underline{y}[2.s_j]$, $j = 1, \ldots, n$. Fisher's exact test is also numerically equivalent to the process-based permutation test of $H_0^{DUP}$ when $(\underline{S}, \underline{T}) \perp\!\!\!\perp \underline{Y}$ and $\underline{Y}[t.i]\text{indep} \sim \text{bin}(1, \pi_t)$; here $H_0^{DUP}$ is equivalent to $\pi_1 = \pi_2$. In fact, in the simulation (scenarios 6 and 7 of Tables 3 and 5, and scenario 6 of Tables 4 and 6), because of the uniform randomization distribution, we were able to use the R code for Fisher's exact test, `fisher.test {stat}`, to compute the exact values of the Fisher randomization and permutation $p$-values. On a related note, we point out that the Neyman randomization test is also available for testing the no-treatment-effect hypothesis $H_0^{RAs}$: $\overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}]$ in $2 \times 2$ tables. This paper's simulation results suggest that when the randomization distribution is uniform as in (7), this Neyman randomization test for $2 \times 2$ tables may be somewhat more powerful than Fisher's exact test.

## ACKNOWLEDGMENTS

Supported in part by NSF Grant SES-1059955.

## REFERENCES


AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York. MR1914507

AGRESTI, A. and FRANKLIN, C. (2007). *Statistics*: *The Art and Science of Learning from Data*. Pearson/Prentice Hall, Upper Saddle River, NJ.

BAILEY, R. A. (1981). A unified approach to design of experiments. *J. Roy. Statist. Soc. Ser. A* **144** 214–223. MR0625801

COPAS, J. B. (1973). Randomization models for the matched and unmatched $2 \times 2$ tables. *Biometrika* **60** 467–476. MR0448746

COX, D. R. (1958a). The interpretation of the effects of non-additivity in the latin square. *Biometrika* **45** 69–73.

COX, D. R. (1958b). *Planning of Experiments*. Wiley, New York. MR1175752

COX, D. R. (2009). Randomization in the design of experiments. *Int. Stat. Rev.* **77** 415–429.

COX, D. R. and REID, N. (2000). *The Theory of the Design of Experiments*. Chapman & Hall/CRC, Boca Raton, FL.

DAVID, H. A. (2008). The beginnings of randomization tests. *Amer. Statist.* **62** 70–72. MR2416900

EDEN, T. and YATES, F. (1933). On the validity of Fisher's $z$ test when applied to an actual example of non-normal data. *J. Agric. Sci.* **23** 6–17.

ERNST, M. D. (2004). Permutation methods: A basis for exact inference. *Statist. Sci.* **19** 676–685. MR2185589

FISHER, R. A. (1935). *The Design of Experiments*. Oliver Boyd, Edinburgh.

GADBURY, G. L. (2001). Randomization inference and bias of standard errors. *Amer. Statist.* **55** 310–313. MR1939365

GREENLAND, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *Amer. Statist.* **45** 248–251.

GREENLAND, S. (2000). Causal analysis in the health sciences. *J. Amer. Statist. Assoc.* **95** 286–289.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York. MR0045368

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50** 946–967. MR0071696

KEMPTHORNE, O. (1977). Why randomize? *J. Statist. Plann. Inference* **1** 1–25. MR0518596

LEHMANN, E. L. (1994). Jerzy Neyman, 1894–1981: A biographical memoir. In *Biographical Memoirs*, *Vol.* 63, *Edited by Office of the Home Secretary*. National Academy of Sciences, Washington, DC.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9. *Roczniki Nauk Rolniczych* Tom **X** [in Polish]; English translation of excerpts by D. M. Dabrowska and T. P. Speed *Statist. Sci.* **5** (1990) 463–472.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive sampling (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **97** 558–625.

NEYMAN, J., IWASKIEWICZ, K. and KOLODZIEJCZYK, S. (1935). Statistical problems in agricultural experimentation (with discussion). *Suppl. J. Roy. Statist. Soc.* **2** 107–180.

PITMAN, E. J. G. (1937). Significance tests which can be applied to samples from any populations. *Suppl. J. Roy. Statist. Soc.* **4** 119–130.

PITMAN, E. J. G. (1938). Significance tests which can be applied to samples from any populations. III. The analysis of variance test. *Biometrika* **29** 322–335.

ROSENBAUM, P. R. (2014). Available at www-stat.wharton.upenn.edu/~rosenbap/downloadTalks.htm.

RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [*Ann. Agric. Sci.* **10** (1923) 1–51]. *Statist. Sci.* **5** 472–480. MR1092987

RUBIN, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *J. Educ. and Behav. Statist.* **29** 343–367.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc*. **100** 322–331. MR2166071

RUBIN, D. B. (2010). Reflections stimulated by the comments of Shadish and West and Thoemmes. *Psychological Methods* **15** 38–46.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

STRAYER, D. L. and JOHNSTON, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science* **12** 462–466.

SUTTER, G., ZYSKIND, G. and KEMPHORNE, O. (1963). Some Aspects of Constrained Randomization ARL Report 63-18, Wright-Patterson AFB, Ohio.

WELCH, B. L. (1937). On the $z$-test in randomized blocks and latin squares. *Biometrika* **29** 21–52.

WELCH, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29** 350–62.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1** 80–83.